

#5



SUBSTITUTE SPECIFICATION

CELL DIFFERENTIATION cDNAs INDUCED BY RETINOIC ACID

5 This application claims the benefit of provisional application Serial No. 60/263,031, filed 18 January 2001.

FIELD OF THE INVENTION

10 The invention relates to cDNAs produced from transcripts induced by retinoic acid and to their use in diagnosis, prognosis, and treatment of cancer and disorders associated with cell differentiation.

BACKGROUND OF THE INVENTION

15 Retinoic acid (RA) is the bioactive metabolite of vitamin A (retinol) which acts on cells to establish or change the pattern of gene activity. RA is well known for its importance in the regulation of cell growth and differentiation (Perez-Castro *et al.* (1989) *Proc Natl Acad Sci* 86:8813-8817; Wei *et al.* (1989) *Mol Endocrinol* 3:454-463), and in the prevention of cancer. In its post-embryonic role, RA has been shown to play a part in limb, spinal cord, lung, and hair cell regeneration. RA may act alone, in association with its receptors, or in combination with melatonin, lipid-soluble vitamins, cytokines and other growth factors, and extracellular matrix molecules such as fibronectin. As an anti-tumor agent, RA is known to induce differentiation or apoptosis (Hansen *et al.* (2000) *Carcinogenesis* 21:1271-1279; Manna and Aggarwal (2000) *Oncogene* 19:2110-2119; and Martinet *et al.* (2000) *Cancer Res* 60:2869-2875).

20 Although several RA-related compounds have demonstrated anti-tumor effects and are in clinical trials for human therapy (De Coster *et al.* (1996) *J Steroid Biochem Mol Biol* 56:133-143; Krekels *et al.* (1996) *Prostate* 29:36-41; and Zhang *et al.* (2000) *J Cell Physiol* 185:1-20), the genes responsible for these effects of RA have only partially been elucidated.

25 Identification of additional cDNAs whose transcripts are induced by RA and participate in the process of cell differentiation satisfies a need in the art by providing new compositions which are useful in the diagnosis, prognosis, and treatment of individuals with cancer or in need of cell or tissue-specific developmental intervention.

SUMMARY OF THE INVENTION

30 The invention provides for a combination comprising a plurality of cDNAs having the nucleic acid sequences of SEQ ID NOs:1-5 that are co-expressed with one or more known genes whose transcripts are induced by retinoic acid or the complements of SEQ ID NOs:1-5. The invention also provides an isolated cDNA having a nucleic acid sequence selected from SEQ ID NOs:1-5 and the complements thereof. In one aspect, the combination is used in the diagnosis, prognosis, and treatment of cancer and disorders associated with cell differentiation. In another aspect, the cDNA is used as a probe, in an expression vector, or in a composition in combination with a labeling moiety.

35 The invention provides a method for using a combination or a cDNA of the invention to screen a

PB-0017 US

plurality of molecules to identify ligands, molecules or compounds, which bind the cDNA. The molecules or compounds are selected from DNA molecules, RNA molecules, peptide nucleic acids (PNAs), mimetics, and proteins. In one embodiment, the combination or the cDNA are attached to a substrate. In one aspect, the

substrate is used to detect gene expression in a diagnosing a cancer or cell differentiation disorder. The method comprises hybridizing the substrate containing the cDNA to a sample under conditions for formation of one or more hybridization complexes, detecting hybridization complex formation; and comparing the amount of complex formation with the amount of complex formation in a non-diseased sample, wherein the altered amount of complex formation indicates the presence of the cancer or cell differentiation disorder.

The invention provides a purified protein encoded by a cDNA of the invention that is co-expressed with one or more known retinoic acid induced genes in a plurality of biological samples. The invention also provides a method for using a protein to screen a plurality of molecules to identify at least one ligand which specifically binds the protein. The molecules are selected from DNA molecules, RNA molecules, peptide nucleic acids, proteins, agonists, antagonists, and antibodies. The invention further provides a method of using a protein to purify a ligand.

The invention provides a method of using a protein to prepare and purify an antibody that specifically binds to the protein of the invention. The invention also provided a purified antibody. The invention further provides a composition comprising a cDNA, a protein or an antibody that specifically binds a protein and a pharmaceutical carrier.

The invention provides a method for using an antibody to detect expression in a sample, the method comprising combining the antibody with a sample under conditions which allow the formation of antibody:protein complexes; and detecting complex formation, wherein complex formation indicates expression of the protein in the sample. In one embodiment, complex formation is compared with standards and is diagnostic of a disorder associated with steroid-responsive tissues or pregnancy. The invention also provides a method for using an antibody to immunopurify a protein comprising attaching the antibody to a substrate; contacting the antibody with solution containing the protein, thereby forming an antibody:protein complex; dissociating the antibody:protein complex; and collecting the purified protein.

BRIEF DESCRIPTION OF THE SEQUENCE LISTING

The Sequence Listing provides exemplary cDNAs comprising the nucleic acid sequences of SEQ ID NOs:1-5. Each sequence is identified by a sequence identification number (SEQ ID NO) and by the Incyte number with which the sequence was first identified.

DESCRIPTION OF THE INVENTION

It must be noted that as used herein and in the appended claims, the singular forms "a", "an", and

PB-0017 US

"the" include the plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a host cell" includes a plurality of such host cells, and a reference to "an antibody" is a reference to one or more antibodies and equivalents thereof known to those skilled in the art.

It is to be understood that this invention is not limited to the particular devices, machines, materials and methods described. Although particular embodiments are presented, equivalent embodiments may be used to practice the invention. The embodiments are provided to illustrate the invention and are not intended to limit the scope of the invention which is limited only by the appended claims.

DEFINITIONS

"Array" refers to an ordered arrangement of at least two cDNAs, proteins, or antibodies on a substrate. At least one of the cDNAs, proteins, or antibodies represents a control or standard, and the other, a cDNA, protein, or antibody of diagnostic or therapeutic interest. The arrangement of two to about 40,000 cDNAs, proteins, or antibodies on the substrate assures that the size and signal intensity of each labeled complex, formed between each cDNA and at least one nucleic acid, or antibody:protein complex, formed between each antibody and at least one protein to which the antibody specifically binds, is individually distinguishable.

"Cancer" refers to any cancer including, but not limited to, an adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, and teratocarcinoma of the adrenal gland, bladder, blood, bone, bone marrow, brain, breast, gastrointestinal tract, heart, kidney, liver, lung, lymph, muscle, nerve, ovary, pancreas, prostate, skin, spleen, stomach, testis, and uterus and particularly cancers or tumors of the bladder, brain, breast, colon, endometrium, ovary, prostate, testicles and uterus.

"cDNA" refers to an isolated polynucleotide. It may be of genomic or synthetic origin, double-stranded or single-stranded, and combined with other cDNAs, vitamins, minerals, carbohydrates, lipids, proteins, or other types of nucleic acids to perform a particular activity or to form a useful composition.

A "combination" refers to at least two and up to 5 cDNAs have nucleic acid sequences selected from SEQ ID NOs:1-5 and their complements as presented in the Sequence Listing.

The "complement" of a nucleic acid molecule of the Sequence Listing refers to a cDNA which is completely complementary over the full length of the sequence and which will hybridize to the nucleic acid molecule under conditions of high stringency.

"Differential expression" refers to an increased, upregulated or present, or decreased, downregulated or absent, gene expression as detected by the absence, presence, or at least two-fold changes in the amount of transcribed messenger RNA or translated protein in a sample.

"Disorders associated with cell differentiation" refers to conditions, diseases and disorders such as amyotrophic lateral sclerosis, Alzheimer disease, amyloidosis, asthma, ataxias, cerebral agenesis, collagen

PB-0017 US

vascular diseases, diabetes, ductus arteriosus, Huntington disease, hypoplastic left heart, psoriasis, retinal disease, rheumatoid arthritis, Smith-Magenis syndrome, and scleroderma.

"Isolated or purified" refers to a cDNA or protein that is removed from its natural environment and that is separated from other components with which it is naturally present.

5 "Labeling moiety" refers to any reporter molecule, visible or radioactive label, than can be attached to or incorporated into a cDNA, protein or antibody. Visible labels include but are not limited to anthocyanins, green fluorescent protein (GFP), β glucuronidase, luciferase, Cy3 and Cy5, and the like. Radioactive markers include radioactive forms of hydrogen, iodine, phosphorous, sulfur, and the like.

10 "Ligand" refers to any agent, molecule, or compound which will bind specifically to a complementary site on a cDNA molecule or polynucleotide, or to an epitope or a protein. Such ligands stabilize or modulate the activity of polynucleotides or proteins and may be composed of inorganic or organic substances including nucleic acids, proteins, carbohydrates, fats, and lipids.

15 A "portion" of a protein refers to that length of amino acid sequence which would retain at least one biological activity, a domain identified by PFAM or PRINTS analysis or an antigenic epitope of the protein identified using Kyte-Doolittle algorithms of the PROTEAN program (DNASTAR, Madison WI).

"Protein" refers to a polypeptide, or a portion thereof, whether naturally occurring, recombinant, or synthetic. An "oligopeptide" is an amino acid sequence from about five residues to about 15 residues that is used as part of a fusion protein to produce an antibody.

20 "Retinoic acid induced gene" refers to a polynucleotide which has been previously identified as useful in the diagnosis, prognosis, or treatment of cancer or disorders associated with cell differentiation. Typically, this means that the known gene is differentially expressed at higher (or lower) levels in tissues from patients with the disorder when compared with normal expression in any tissue. The "known retinoic acid induced genes" as identified by library subtraction methodology are: P450 retinoic acid induced 1 and 2 (RAI-1/2), retinoic acid hydroxylase (CYP26), laminin beta-2 chain (laminin β -2), retinol-binding-protein receptor p63 (RBPR-p63), P97, lamin A, beta-2-microglobulin (β -2-m), and amyloid precursor like protein 2 (APLP2). The known cell differentiation genes are collagens, notch 3, platelet derived growth factor (PDGF), fibulin-2, fibrillin, insulin growth factor-1 (IGF-I), cAMP-dependent protein kinase regulatory subunit RI beta (RPK-RI), SWAP, EMP, YPT3/rab11, ERK-1, and β 4 integrin.

30 "Sample" is used in its broadest sense as containing nucleic acids, proteins, antibodies, and the like. A sample may comprise a bodily fluid; the soluble fraction of a cell preparation, or an aliquot of media in which cells were grown; a chromosome, an organelle, or membrane isolated or extracted from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; a biopsy, a cell; a tissue; a tissue print; a fingerprint, buccal cells, skin, or hair; and the like.

"Specific binding" refers to a special and precise interaction between two molecules which is

PB-0017 US

dependent upon their structure, particularly their molecular side groups. For example, the intercalation of a regulatory protein into the major groove of a DNA molecule, the hydrogen bonding along the backbone between two single stranded nucleic acids, or the binding between an epitope of a protein and an agonist, antagonist, or antibody.

"Substrate" refers to any rigid or semi-rigid support to which cDNAs or proteins are bound and includes membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, capillaries or other tubing, plates, polymers, and microparticles with a variety of surface forms including wells, trenches, pins, channels and pores.

A "variant" refers to a cDNA or protein whose sequence diverges from about 5% to about 30% from the nucleic acid or amino acid sequences of the Sequence Listing.

THE INVENTION

The present invention utilizes a method for identifying cDNAs and proteins that are associated with a signaling or regulatory pathway specific disease, subcellular compartment, cell type, tissue, or species. In particular, the method identifies cDNAs derived from transcripts of genes which are induced by retinoic acid and that are co-expressed with genes known to be involved in cell differentiation. These genes and cDNAs are useful in diagnosis, prognosis, treatment, and evaluation of therapies for cancers and disorders associated with cell and tissue development and differentiation.

Expression levels of genes in two cell lines, a bone marrow-derived cell line and a fibroblast-derived cell line, before and after treatment with retinoic acid (RA) were compared using library subtraction. The bone-marrow cDNA libraries were produced from cells derived from a metastatic bone marrow neuroblastoma removed from a 4-year old Caucasian female. The samples were either untreated or treated with 10 μ M RA and cultured for four days. The fibroblast libraries were produced from cells derived from fibroblast tissue removed from the breast of a 31-year old Caucasian female. The samples were either untreated or treated with 1 μ M RA for 20 hours.

The genes and cDNAs identified by library subtraction methodology and expressed at high levels in the retinoic acid-treated libraries and not detected in the untreated libraries include:

Gene/cDNA*Known association with RA

RAI-1/2	Yes
SEQ ID NO:1 (Incyte 238040)	No previous report; uncharacterized cDNA
SEQ ID NO:2 (Incyte 411448)	No previous report; uncharacterized cDNA
CYP26	Yes
Laminin β -2	Yes
RBPR-p63	Yes
SEQ ID NO:5 (Incyte 254749)	No previous report; tumor suppressor
SEQ ID NO:3 (Incyte 454163)	No previous report; uncharacterized cDNA
P97	Yes
SEQ ID NO:4 (Incyte 988966)	No previous report; uncharacterized cDNA

PB-0017 US

Lamin A

Yes

 β -2-m

Yes

APLP2Yes

5 *These genes are ordered by expression level with the highest expression occurring in RAI-1/2.

The cDNAs shown above by SEQ ID NO and Incyte number were further analyzed using the guilt by association method (Walker and Volkmuth (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res 9:1198-1203). The method provides for the identification of cDNAs that are expressed in a plurality of libraries. The cDNAs include genes of
 10 known or unknown function which are expressed in a specific signaling pathway, disease process, subcellular compartment, cell type, tissue, or species. The expression patterns of genes with known function are compared with those of cDNAs with unknown function to determine whether a specified co-expression probability threshold is met. Through this comparison, a subset of the cDNAs having a high co-expression probability with the known genes can be identified. The high co-expression probability
 15 correlates with a particular co-expression probability threshold which is preferably less than 0.001 and more preferably less than 0.00001.

The cDNAs originate from cDNA libraries derived from a variety of sources including, but not limited to, eukaryotes such as human, mouse, rat, dog, monkey, plant, and yeast; prokaryotes such as
 20 bacteria; and viruses. These cDNAs can also be selected from a variety of sequence types including, but not limited to, expressed sequence tags (ESTs), assembled polynucleotides, full length gene coding regions, promoters, introns, enhancers, 5' untranslated regions, and 3' untranslated regions. To have statistically significant analytical results, the cDNAs are expressed in at least five cDNA libraries.

The 1176 cDNA libraries used in the co-expression analysis of the present invention can be obtained from adrenal gland, biliary tract, bladder, blood cells, blood vessels, bone marrow, brain, bronchus,
 25 cartilage, chromaffin system, colon, connective tissue, cultured cells, embryonic stem cells, endocrine glands, epithelium, esophagus, fetus, ganglia, heart, hypothalamus, immune system, intestine, islets of Langerhans, kidney, larynx, liver, lung, lymph, muscles, neurons, ovary, pancreas, penis, peripheral nervous system, phagocytes, pituitary, placenta, pleura, prostate, salivary glands, seminal vesicles, skeleton, spleen, stomach, testis, thymus, tongue, ureter, uterus, and the like. The number of cDNA libraries selected can
 30 range from as few as 5 to greater than 10,000 and preferably, the number of the cDNA libraries is greater than 500.

In a preferred embodiment, the cDNAs are assembled from related sequences, such as sequence fragments derived from a single transcript. Assembly of the polynucleotide can be performed using sequences of various types including, but not limited to, ESTs, extension of the EST, shotgun sequences
 35 from a cloned insert, or full length cDNAs. In a most preferred embodiment, the cDNAs are derived

PB-0017 US

from human sequences that have been assembled using the algorithm disclosed in USSN 9,276,534, filed March 25, 1999, incorporated herein by reference.

Experimentally, differential expression of the polynucleotides can be evaluated by methods including, but not limited to, differential display by spatial immobilization or by gel electrophoresis, genome mismatch scanning, representational difference analysis, and transcript imaging. Additionally, differential expression can be assessed by microarray technology. These methods may be used alone or in combination.

Known retinoic acid-induced genes can be selected based on the use of the genes as diagnostic or prognostic markers or as therapeutic targets for cancer or disorders associated with cell differentiation.

Preferably, the cell differentiation genes induced by retinoic acid are RAI-1/2, CYP26, laminin β -2, RBPR-p63, P97, lamin A, β -2-m, APLP2, collagens, notch 3, PDGF, fibulin-2, fibrillin, IGF-I, RPK RI, SWAP, EMP, YPT3/rab11, ERK-1, and β 4 integrin.

The procedure for identifying novel cDNAs that exhibit a statistically significant co-expression pattern with known retinoic acid induced genes is as follows. First, the presence or absence of a gene sequence in a cDNA library is defined: a gene is present in a cDNA library when at least one cDNA fragment corresponding to that gene is detected in a cDNA sample taken from the library, and a gene is absent from a library when no corresponding cDNA fragment is detected in the sample.

Second, the significance of gene co-expression is evaluated using a probability method to measure a due-to-chance probability of the co-expression. The probability method can be the Fisher exact test, the chi-squared test, or the kappa test. These tests and examples of their applications are well known in the art and can be found in standard statistics texts (Agresti (1990) Categorical Data Analysis, John Wiley & Sons, New York NY; Rice (1988) Mathematical Statistics and Data Analysis, Duxbury Press, Pacific Grove CA). A Bonferroni correction (Rice, supra, p. 384) can also be applied in combination with one of the probability methods for correcting statistical results of one gene versus multiple other genes. In a preferred embodiment, the due-to-chance probability is measured by a Fisher exact test, and the threshold of the due-to-chance probability is set preferably to less than 0.001, more preferably to less than 0.00001.

To determine whether two genes, A and B, have similar co-expression patterns, occurrence data vectors can be generated as illustrated in Table 1. The presence of a gene occurring at least once in a library is indicated by a one, and its absence from the library, by a zero.

Table 1. Occurrence Data for Genes A and B

	Library 1	Library 2	Library 3	...	Library N
Gene A	1	1	0	...	0
Gene B	1	0	1	...	0

PB-0017 US

For a given pair of genes, the occurrence data in Table 1 can be summarized in a 2 x 2 contingency table.

Table 2. Contingency Table for Co-occurrences of Genes A and B

	Gene A Present	Gene A Absent	Total
Gene B Present	8	2	10
Gene B Absent	2	18	20
Total	10	20	30

Table 2 presents co-occurrence data for gene A and gene B in a total of 30 libraries. Both gene A and gene B occur 10 times in the libraries. Table 2 summarizes and presents: 1) the number of times gene A and B are both present in a library; 2) the number of times gene A and B are both absent in a library; 3) the number of times gene A is present, and gene B is absent; and 4) the number of times gene B is present, and gene A is absent. The upper left entry is the number of times the two genes co-occur in a library, and the middle right entry is the number of times neither gene occurs in a library. The off diagonal entries are the number of times one gene occurs, and the other does not. Both A and B are present eight times and absent 18 times. Gene A is present, and gene B is absent, two times; and gene B is present, and gene A is absent, two times. The probability ("p-value") that the above association occurs due to chance as calculated using a Fisher exact test is 0.0003. Associations are generally considered significant if a p-value is less than 0.01 or 1.0e-2 (Agresti, supra; Rice, supra).

This method of estimating the probability for coexpression of two genes makes several assumptions. The method assumes that the libraries are independent and are identically sampled. However, in practical situations, the selected cDNA libraries are not entirely independent, because more than one library may be obtained from a single subject or tissue. Nor are they entirely identically sampled, because different numbers of cDNAs may be sequenced from each library. The number of cDNAs sequenced typically ranges from 5,000 to 10,000 cDNAs per library. In addition, because a Fisher exact coexpression probability is calculated for each gene versus 37,071 other assembled genes that occur in at least five libraries, a Bonferroni correction for multiple statistical tests is used.

Using the method of the present invention, we have identified cDNAs that exhibit significant association or co-expression probability with known genes that are specific to cancer or disorders associated with cell differentiation. The results presented in Example VI show the direct or indirect associations among expression of novel cDNAs and known retinoic acid induced genes and cell differentiation genes. These genes are RAI1, CYP26, laminin β -2, RBPR-p63, P97, lamin A, β -2-m, APLP2, collagens, notch 3, PDGF, fibulin-2, fibrillin, IGF-I, RPK-RI, SWAP, EMP, YPT3/rab11, ERK-1, and β 4 integrin. Therefore, the five novel cDNAs, SEQ ID NOs1-5 of the Sequence Listing, are useful as surrogate markers for the co-expressed genes in diagnosis, prognosis, or treatment of cancer and

PB-0017 US

disorders associated with cell differentiation. Further, the proteins or peptides expressed from the novel cDNAs are either potential therapeutics or targets for the identification or development of therapeutics.

Therefore, in one embodiment, the present invention encompasses a combination comprising a plurality of cDNAs having the nucleic acid sequences of SEQ ID NOs:1-5 or the complements thereof.

These five cDNAs are shown by the method of the present invention to have significant co-expression with known retinoic acid-induced cell differentiation genes. The invention also provides a cDNA, its complement, and a probe comprising the cDNA selected from SEQ ID NOs:1-5. Variants typically have at least about 70%, more preferably at least about 85%, and most preferably at least about 95% nucleic acid sequence identity to at least one of these sequences.

The cDNA or the encoded protein may be used to search against the GenBank primate (pri), rodent (rod), mammalian (mam), vertebrate (vrtp), and eukaryote (eukp) databases, SwissProt, BLOCKS (Bairoch *et al.* (1997) *Nucleic Acids Res* 25:217-221), PFAM, and other databases that contain previously identified and annotated motifs, sequences, and gene functions. Methods that search for primary sequence patterns with secondary structure gap penalties (Smith *et al.* (1992) *Protein Engineering* 5:35-51) as well as algorithms such as Basic Local Alignment Search Tool (BLAST; Altschul (1993) *J Mol Evol* 36:290-300; Altschul *et al.* (1990) *J Mol Biol* 215:403-410), BLOCKS (Henikoff and Henikoff (1991) *Nucleic Acids Res* 19:6565-6572), Hidden Markov Models (HMM; Eddy (1996) *Cur Opin Str Biol* 6:361-365; Sonnhammer *et al.* (1997) *Proteins* 28:405-420), and the like, can be used to manipulate and analyze nucleotide and amino acid sequences. These databases, algorithms and other methods are well known in the art and are described in Ausubel *et al.* (1997; *Short Protocols in Molecular Biology*, John Wiley & Sons, New York NY, unit 7.7) and in Meyers (1995; *Molecular Biology and Biotechnology*, Wiley VCH, New York NY, p 856-853).

Also encompassed by the invention are polynucleotides that are capable of hybridizing to SEQ ID NOs:1-5, and fragments thereof under stringent conditions. Stringent conditions can be defined by salt concentration, temperature, and other chemicals and conditions well known in the art. Conditions can be selected, for example, by varying the concentrations of salt in the prehybridization, hybridization, and wash solutions or by varying the hybridization and wash temperatures. With some substrates, the temperature can be decreased by adding formamide to the prehybridization and hybridization solutions.

Hybridization can be performed at low stringency, with buffers such as 5xSSC (saline sodium citrate) with 1% sodium dodecyl sulfate (SDS) at 60°C, which permits complex formation between two nucleic acid sequences that contain some mismatches. Subsequent washes are performed at higher stringency with buffers such as 0.2xSSC with 0.1% SDS at either 45°C (medium stringency) or 68°C (high stringency), to maintain hybridization of only those complexes that contain completely complementary sequences. Background signals can be reduced by the use of detergents such as SDS,

PB-0017 US

sarcosyl, or TRITON X-100 (Sigma-Aldrich, St. Louis MO), and/or a blocking agent, such as salmon sperm DNA. Hybridization methods are described in detail in Ausubel (supra, units 2.8-2.11, 3.18-3.19 and 4-6-4.9) and Sambrook et al. (1989; Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview NY)

5 A cDNA can be extended utilizing a partial nucleotide sequence and employing various PCR-based methods known in the art to detect upstream sequences such as promoters and other regulatory elements. (See, e.g., Dieffenbach and Dveksler (1995) PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview NY). Additionally, an XL-PCR kit (Applied Biosystems (ABI), Foster City CA), nested primers, and commercially available cDNA libraries (Invitrogen, Carlsbad CA) or genomic libraries
10 (Clontech, Palo Alto CA) may be used to extend the sequence. For all PCR-based methods, primers may be designed using commercially available software (LASERGENE software, DNASTAR) or another program, to be about 15 to 30 nucleotides in length, to have a GC content of about 50%, and to form a hybridization complex at temperatures of about 68°C to 72°C.

In another aspect of the invention, the cDNA can be cloned into a recombinant vector that directs
15 the expression of the protein, or structural or functional portions thereof, in host cells. Due to the inherent degeneracy of the genetic code, other DNA sequences which encode substantially the same or a functionally equivalent amino acid sequence may be produced and used to express the protein encoded by the cDNA. The nucleotide sequences of the present invention can be engineered using methods generally known in the art in order to alter the nucleotide sequences for a variety of purposes including, but not
20 limited to, modification of the cloning, processing, and/or expression of the gene product. DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides may be used to engineer the nucleotide sequences. For example, oligonucleotide-mediated site-directed mutagenesis may be used to introduce mutations that create new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, and so forth.

25 In order to express a biologically active protein, the cDNA or derivatives thereof, may be inserted into an expression vector, i.e., a vector which contains the elements for transcriptional and translational control of the inserted coding sequence in a particular host. These elements include regulatory sequences, such as enhancers, constitutive and inducible promoters, and 5' and 3' untranslated regions. Methods which are well known to those skilled in the art may be used to construct such expression vectors. These
30 methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination (Sambrook, supra; Ausubel, supra).

A variety of expression vector/host cell systems may be utilized to express the cDNA. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid expression vectors; yeast transformed with yeast expression vectors;

PB-0017 US

insect cell systems infected with baculovirus vectors; plant cell systems transformed with viral or bacterial expression vectors; or animal cell systems. For long term production of recombinant proteins in mammalian systems, stable expression in cell lines is preferred. For example, the cDNA can be transformed into cell lines using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable or visible marker gene on the same or on a separate vector. The invention is not to be limited by the vector or host cell employed.

In general, host cells that contain the cDNA and that express the protein may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations, PCR amplification, and protein bioassay or immunoassay techniques which include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or amino acid sequences. Immunological methods for detecting and measuring the expression of the protein using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS).

Host cells transformed with the cDNA may be cultured under conditions for the expression and recovery of the protein from cell culture. The protein produced by a transgenic cell may be secreted or retained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing the cDNA may be designed to contain signal sequences which direct secretion of the protein through a prokaryotic or eukaryotic cell membrane.

In addition, a host cell strain may be chosen for its ability to modulate expression of the inserted sequences or to process the expressed protein in the desired fashion. Such modifications of the protein include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation. Post-translational processing which cleaves a "prepro" form of the protein may also be used to specify protein targeting, folding, and/or activity. Different host cells which have specific cellular machinery and characteristic mechanisms for post-translational activities (e.g., CHO, HeLa, MDCK, HEK293, and WI38) are available from the ATCC (Manassas VA) and may be chosen to ensure the correct modification and processing of the expressed protein.

In another embodiment of the invention, natural, modified, or recombinant nucleic acid sequences are ligated to a heterologous sequence resulting in translation of a fusion protein containing heterologous protein moieties in any of the aforementioned host systems. Such heterologous protein moieties facilitate purification of fusion proteins using commercially available affinity matrices. Such moieties include, but are not limited to, glutathione S-transferase, maltose binding protein, thioredoxin, calmodulin binding peptide, 6-His, FLAG, c-myc, hemagglutinin, and monoclonal antibody epitopes.

In another embodiment, the cDNAs, wholly or in part, are synthesized using chemical or

PB-0017 US

enzymatic methods well known in the art (Caruthers *et al.* (1980) Nucl Acids Symp Ser (7) 215-233; Ausubel, *supra*). For example, peptide synthesis can be performed using various solid-phase techniques (Roberge *et al.* (1995) Science 269:202-204), and machines such as the ABI 431A peptide synthesizer (ABI) can be used to automate synthesis. If desired, the amino acid sequence may be altered during synthesis and/or combined with sequences from other proteins to produce a variant.

SCREENING, DIAGNOSTICS AND THERAPEUTICS

The cDNAs can be used as surrogate markers in diagnosis, prognosis, treatment, and selection and evaluation of therapies for cancer and disorders associated with cell differentiation, both as defined herein.

The cDNAs may be used to screen a plurality of molecules for specific binding affinity. The assay can be used to screen a plurality of DNA molecules, RNA molecules, peptide nucleic acids, peptides, ribozymes, antibodies, agonists, antagonists, immunoglobulins, inhibitors, proteins including transcription factors, enhancers, repressors, and drugs and the like which regulate the activity of the polynucleotide in the biological system. The assay involves providing a plurality of molecules, contacting the cDNAs of the combination with the plurality of molecules under conditions suitable to allow specific binding, and detecting specific binding to identify at least one molecule which specifically binds the cDNA.

Similarly the proteins or portions thereof may be used to screen libraries of molecules or compounds in any of a variety of screening assays. The portion of a protein employed in such screening may be free in solution, affixed to an abiotic or biotic substrate (e.g. borne on a cell surface), or located intracellularly. Specific binding between the protein and the molecule may be measured. The assay can be used to screen a plurality of DNA molecules, RNA molecules, PNAs, peptides, mimetics, ribozymes, antibodies, agonists, antagonists, immunoglobulins, inhibitors, peptides, polypeptides, drugs and the like, which specifically bind the protein. One method for high throughput screening using very small assay volumes and very small amounts of test compound is described in USPN 5,876,946, incorporated herein by reference, which screens large numbers of molecules for enzyme inhibition or receptor binding.

In one preferred embodiment, the cDNAs are used for diagnostic purposes to determine the absence, presence, or altered--increased or decreased compared to a normal standard-- expression of the gene. The polynucleotide consists of complementary RNA and DNA molecules, branched nucleic acids, and/or PNAs. In one alternative, the polynucleotides are used to detect and quantify gene expression in samples in which expression of the cDNA is correlated with disease. In another alternative, the cDNA can be used to detect genetic polymorphisms associated with a disease. These polymorphisms may be detected in the transcript cDNA.

The specificity of the probe is determined by whether it is made from a unique region, a regulatory region, or from a conserved motif. Both probe specificity and the stringency of diagnostic

PB-0017 US

hybridization or amplification (maximal, high, intermediate, or low) will determine whether the probe identifies only naturally occurring, exactly complementary sequences, allelic variants, or related sequences. Probes designed to detect related sequences should preferably have at least 50% sequence identity to any of the polynucleotides encoding the protein.

5 Methods for producing hybridization probes include the cloning of nucleic acid sequences into vectors for the production of RNA probes. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by adding RNA polymerases and labeled nucleotides. Hybridization probes may incorporate nucleotides labeled by a variety of reporter groups including, but not limited to, radionuclides such as ^{32}P or ^{35}S , enzymatic labels such as alkaline phosphatase coupled to the
10 probe via avidin/biotin coupling systems, fluorescent labels, and the like. The labeled cDNAs may be used in Southern or northern analysis, dot blot, or other membrane-based technologies; in PCR technologies; and in microarrays utilizing samples from subjects to detect altered protein expression.

 The cDNA can be labeled by standard methods and added to a sample from a subject under conditions for the formation and detection of hybridization complexes. After incubation the sample is
15 washed, and the signal associated with hybrid complex formation is quantitated and compared with a standard value. Standard values are derived from any control sample, typically one that is free of the suspect disease. If the amount of signal in the subject sample is altered in comparison to the standard value, then the presence of altered levels of expression in the sample indicates the presence of the disease. Qualitative and quantitative methods for comparing the hybridization complexes formed in
20 subject samples with previously established standards are well known in the art.

 Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual subject. Once the presence of disease is established and a treatment protocol is initiated, hybridization or amplification assays can be repeated on a regular basis to determine if the level of expression in the patient begins to
25 approximate that which is observed in a healthy subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to many years.

 The cDNAs may be used for the diagnosis of a variety of cancers and disorders associated with cell differentiation.

 The cDNAs may be used on a substrate such as microarray to monitor the expression patterns.
30 Arrays incorporating cDNAs, proteins, or antibodies may be prepared and analyzed using methods well known in the art. Oligonucleotides or cDNAs may be used as hybridization probes or targets to monitor the expression level of large numbers of genes simultaneously or to identify genetic variants, mutations, and single nucleotide polymorphisms. Proteins may be used to identify ligands, to investigate protein:protein interactions, or to produce a proteomic profile of gene expression (i.e., to detect and

PB-0017 US

quantify expression of a protein in a sample). Antibodies may be also be used produce a proteomic profile of gene expression. Such arrays may be used to determine gene function; to understand the genetic basis of a condition, disease, or disorder; to diagnose a condition, disease, or disorder; and to develop and monitor the activities of therapeutic agents. (See, e.g., Brennan et al. (1995) USPN 5,474,796; Schena et al. (1996) Proc Natl Acad Sci 93:10614-10619; Heller et al. (1997) Proc Natl Acad Sci 94:2150-2155; Heller et al. (1997) USPN 5,605,662; and deWildt et al. (2000) Nature Biotechnol 18:989-994.)

In another embodiment, antibodies or Fabs comprising an antigen binding site that specifically binds the protein may be used for the diagnosis of diseases characterized by the over-or-under expression of the protein. A variety of protocols for measuring protein expression, including ELISAs, RIAs, and FACS, are well known in the art and provide a basis for diagnosing altered or abnormal levels of expression. Standard values for protein expression are established by combining samples taken from healthy subjects, preferably human, with antibody to the protein under conditions for complex formation. The amount of complex formation may be quantitated by various methods, preferably by photometric means. Quantities of the protein expressed in disease samples are compared with standard values.

Deviation between standard and subject values establishes the parameters for diagnosing or monitoring disease. Alternatively, one may use competitive drug screening assays in which neutralizing antibodies capable of binding specifically with the protein compete with a test compound. Antibodies can be used to detect the presence of any peptide which shares one or more antigenic determinants with the protein. In one aspect, the antibodies of the present invention can be used for treatment or monitoring therapeutic treatment for cancers and disorders associated with cell and tissue development and differentiation.

In another aspect, the cDNA, or its complement, may be used therapeutically for the purpose of expressing mRNA and protein, or conversely to block transcription or translation of the mRNA. Expression vectors may be constructed using elements from retroviruses, adenoviruses, herpes or vaccinia viruses, or bacterial plasmids, and the like. These vectors may be used for delivery of nucleotide sequences to a particular target organ, tissue, or cell population. Methods well known to those skilled in the art can be used to construct vectors to express nucleic acid sequences or their complements. (See, e.g., Maulik et al. (1997) Molecular Biotechnology, Therapeutic Applications and Strategies, Wiley-Liss, New York NY.) Alternatively, the cDNA or its complement, may be used for somatic cell or stem cell gene therapy. Vectors may be introduced in vivo, in vitro, and ex vivo. For ex vivo therapy, vectors are introduced into stem cells taken from the subject, and the resulting transgenic cells are clonally propagated for autologous transplant back into that same subject. Delivery of the cDNA by transfection, liposome injections, or polycationic amino polymers may be achieved using methods which are well known in the art. (See, e.g., Goldman et al. (1997) Nature Biotechnol 15:462-466.) Additionally, endogenous gene expression may be inactivated using homologous recombination methods which insert an inactive gene

PB-0017 US

sequence into the coding region or other targeted region of the cDNA. (See, e.g. Thomas et al. (1987) Cell 51:503-512.)

Vectors containing the cDNA can be transformed into a cell or tissue to express a missing protein or to replace a nonfunctional protein. Similarly a vector constructed to express the complement of the cDNA can be transformed into a cell to downregulate the protein expression. Complementary or antisense sequences may consist of an oligonucleotide derived from the transcription initiation site; nucleotides between about positions -10 and +10 from the ATG are preferred. Similarly, inhibition can be achieved using triple helix base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature. (See, e.g., Gee et al. In: Huber and Carr (1994) Molecular and Immunologic Approaches, Futura Publishing, Mt. Kisco NY, pp. 163-177.)

Ribozymes, enzymatic RNA molecules, may also be used to catalyze the cleavage of mRNA and decrease the levels of particular mRNAs, such as those comprising the cDNAs of the invention. (See, e.g., Rossi (1994) Current Biology 4: 469-471.) Ribozymes may cleave mRNA at specific cleavage sites. Alternatively, ribozymes may cleave mRNAs at locations dictated by flanking regions that form complementary base pairs with the target mRNA. The construction and production of ribozymes is well known in the art and is described in Meyers (supra).

RNA molecules may be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at the 5' and/or 3' ends of the molecule, or the use of phosphorothioate or 2' O-methyl rather than phosphodiester linkages within the backbone of the molecule. Alternatively, nontraditional bases such as inosine, queosine, and wybutosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytidine, guanine, thymine, and uridine which are not as easily recognized by endogenous endonucleases, may be included.

Further, an antagonist, or an antibody that binds specifically to the protein may be administered to a subject to treat a cancer or a disorder associated with cell differentiation. The antagonist, antibody, or fragment may be used directly to inhibit the activity of the protein or indirectly to deliver a therapeutic agent to cells or tissues which express the protein. The therapeutic agent may be a cytotoxic agent selected from a group including, but not limited to, abrin, ricin, doxorubicin, daunorubicin, taxol, ethidium bromide, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicine, dihydroxy anthracin dione, actinomycin D, diphtheria toxin, Pseudomonas exotoxin A and 40, radioisotopes, and glucocorticoid.

Antibodies to the protein may be produced using methods that are well known in the art. Such antibodies may include, but are not limited to, polyclonal, monoclonal, chimeric, and single chain antibodies, Fab fragments, and fragments produced by a Fab expression library. Neutralizing antibodies, such as

PB-0017 US

those which inhibit dimer formation, are especially preferred for therapeutic use. Monoclonal antibodies to the protein may be prepared using any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma, the human B-cell hybridoma, and the EBV-hybridoma techniques. In addition, techniques developed for the production of chimeric antibodies can be used. (See, e.g., Pound (1998) Immunochemical Protocols, Methods Mol Biol Vol. 80). Alternatively, techniques described for the production of single chain antibodies may be employed. Fabs which contain specific binding sites for the protein may also be generated. Various immunoassays may be used to identify antibodies having the desired specificity. Numerous protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies with established specificities are well known in the art.

Yet further, an agonist of the protein may be administered to a subject to treat or prevent a disease associated with decreased expression, longevity or activity of the protein.

An additional aspect of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic applications discussed above. Such pharmaceutical compositions may consist of the protein or antibodies, mimetics, agonists, antagonists, or inhibitors of the protein. The compositions may be administered alone or in combination with at least one other agent, such as a stabilizing compound, which may be administered in any sterile, biocompatible pharmaceutical carrier including, but not limited to, saline, buffered saline, dextrose, and water. The compositions may be administered to a subject alone or in combination with other agents, drugs, or hormones.

The pharmaceutical compositions utilized in this invention may be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

In addition to the active ingredients, these pharmaceutical compositions may contain pharmaceutically-acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Mack Publishing, Easton PA).

For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays or in animal models such as mice, rats, rabbits, dogs, or pigs. An animal model may also be used to determine the concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient which ameliorates the

PB-0017 US

symptoms or condition. Therapeutic efficacy and toxicity may be determined by standard pharmaceutical procedures in cell cultures or with experimental animals, such as by calculating and contrasting the ED₅₀ (the dose therapeutically effective in 50% of the population) and LD₅₀ (the dose lethal to 50% of the population) statistics. Any of the therapeutic compositions described above may be applied to any subject in need of such therapy, including, but not limited to, mammals such as dogs, cats, cows, horses, rabbits, monkeys, and most preferably, humans.

Stem Cells and Their Use

SEQ ID NOs:1-5 may be useful in the differentiation of stem cells. Eukaryotic stem cells are able to differentiate into the multiple cell types of various tissues and organs and to play roles in embryogenesis and adult tissue regeneration (Gearhart (1998) Science 282:1061-1062; Watt and Hogan (2000) Science 287:1427-1430). Depending on their source and developmental stage, stem cells may be totipotent with the potential to create every cell type in an organism and to generate a new organism, pluripotent with the potential to give rise to most cell types and tissues, but not a whole organism; or multipotent cells with the potential to differentiate into a limited number of cell types. Stem cells may be transfected with polynucleotides which may be transiently expressed or may be integrated within the cell as transgenes.

Embryonic stem (ES) cell lines are derived from the inner cell masses of human blastocysts and are pluripotent (Thomson *et al.* (1998) Science 282:1145-1147). They have normal karyotypes and express high levels of telomerase which prevents senescence and allows the cells to replicate indefinitely. ES cells produce derivatives that give rise to embryonic epidermal, mesodermal and endodermal cells. Embryonic germ (EG) cell lines, which are produced from primordial germ cells isolated from gonadal ridges and mesenteries, also show stem cell behavior (Shamblott *et al.* (1998) Proc Natl Acad Sci 95:13726-13731). EG cells have normal karyotypes and appear to be pluripotent.

Organ-specific adult stem cells differentiate into the cell types of the tissues from which they were isolated. They maintain their original tissues by replacing cells destroyed from disease or injury. Adult stem cells are multipotent and under proper stimulation can be used to generate cell types of various other tissues (Vogel (2000) Science 287:1418-1419). Hematopoietic stem cells from bone marrow provide not only blood and immune cells, but can also be induced to transdifferentiate to form brain, liver, heart, skeletal muscle and smooth muscle cells. Similarly mesenchymal stem cells can be used to produce bone marrow, cartilage, muscle cells, and some neuron-like cells, and stem cells from muscle have the ability to differentiate into muscle and blood cells (Jackson *et al.* (1999) Proc Natl Acad Sci 96:14482-14486). Neural stem cells, which produce neurons and glia, may also be induced to differentiate into heart, muscle, liver, intestine, and blood cells (Kuhn and Svendsen (1999) BioEssays 21:625-630); Clarke *et al.* (2000) Science 288:1660-1663; Gage (2000) Science 287:1433-1438; and Galli *et al.* (2000) Nature

PB-0017 US

Neurosci 3:986-991).

Neural stem cells may be used to treat neurological disorders such as Alzheimer disease, Parkinson disease, and multiple sclerosis and to repair tissue damaged by strokes and spinal cord injuries. Hematopoietic stem cells may be used to restore immune function in immunodeficient patients or to treat autoimmune disorders by replacing autoreactive immune cells with normal cells to treat diseases such as multiple sclerosis, scleroderma, rheumatoid arthritis, and systemic lupus erythematosus. Mesenchymal stem cells may be used to repair tendons or to regenerate cartilage to treat arthritis. Liver stem cells may be used to repair liver damage. Pancreatic stem cells may be used to replace islet cells to treat diabetes. Muscle stem cells may be used to regenerate muscle to treat muscular dystrophies. (See, for example, 5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185
9190
9195
9200
9205
9210
9215
9220
9225
9230
9235
9240
9245
9250
9255
9260
9265
9270
9275
9280
9285
9290
9295
9300
9305
9310
9315
9320
9325
9330
9335
9340
9345
9350
9355
9360
9365
9370
9375
9380
9385
9390
9395
9400
9405
9410
9415
9420
9425
9430
9435
9440
9445
9450
9455
9460
9465
9470
9475
9480
9485
9490
9495
9500
9505
9510
9515
9520
9525
9530
9535
9540
9545
9550
9555
9560
9565
9570
9575
9580
9585
9590
9595
9600
9605
9610
9615
9620
9625
9630
9635
9640
9645
9650
9655
9660
9665
9670
9675
9680
9685
9690
9695
9700
9705
9710
9715
9720
9725
9730
9735
9740
9745
9750
9755
9760
9765
9770
9775
9780
9785
9790
9795
9800
9805
9810
9815
9820
9825
9830
9835
9840
9845
9850
9855
9860
9865
9870
9875
9880
9885
9890
9895
9900
9905
9910
9915
9920
9925
9930
9935
9940
9945
9950
9955
9960
9965
9970
9975
9980
9985
9990
9995
10000
10005
10010
10015
10020
10025
10030
10035
10040
10045
10050
10055
10060
10065
10070
10075
10080
10085
10090
10095
10100
10105
10110
10115
10120
10125
10130
10135
10140
10145
10150
10155
10160
10165
10170
10175
10180
101

PB-0017 US

plasmid system (Invitrogen). The cDNAs were fractionated on a SEPHAROSE CL4B column (Amersham Pharmacia Biotech (APB), Piscataway NJ), and those cDNAs exceeding 400 bp were ligated into pINCY plasmid (Incyte Genomics, Palo Alto CA) and transformed into DH5 α competent cells (Invitrogen).

Bone marrow cDNA synthesis was initiated using a NotI-anchored oligo d(T) primer. Double-stranded cDNA was blunted, ligated to EcoRI adaptors, digested with NotI, size-selected, and cloned into the NotI and EcoRI sites of the pINCY vector (Incyte Genomics) and transformed into competent cells.

II Isolation and Sequencing of cDNA Clones

Plasmid DNA was released from the cells and purified using the REAL PREP 96 plasmid kit (Qiagen). The recommended protocol was employed except for the following changes: 1) the bacteria were cultured in 1 ml of sterile TERRIFIC BROTH (BD Biosciences, Sparks MD) with carbenicillin at 25 mg/l and glycerol at 0.4%; 2) the cultures were incubated for 19 hours after the wells were inoculation and then lysed with 0.3 ml of lysis buffer; 3) following isopropanol precipitation, the DNA pellet was resuspended in 0.1 ml of distilled water. After the last step in the protocol, samples were transferred to a 96-well block for storage at 4° C.

The cDNAs were prepared using a MICROLAB 2200 system (Hamilton, Reno NV) in combination with DNA ENGINE thermal cyclers (MJ Research, Watertown MA). The cDNAs were sequenced by the method of Sanger and Coulson (1975; J Mol Biol 94:441f) using ABI PRISM 377 DNA sequencing systems (ABI). Most of the sequences were sequenced using standard ABI protocols and kits (ABI) at solution volumes of 0.25x - 1.0x. In the alternative, some of the sequences were sequenced using solutions and dyes from APB.

III Selection, Assembly, and Characterization of Sequences

The sequences used for co-expression analysis were assembled from EST sequences, 5' and 3' long read sequences, and full length coding sequences. Selected assembled sequences were expressed in at least three cDNA libraries.

The assembly process is described as follows. EST sequence chromatograms were processed and verified. Quality scores were obtained using PHRED (Ewing *et al.* (1998) Genome Res 8:175-185; Ewing and Green (1998) Genome Res 8:186-194), and edited sequences were loaded into a relational database management system (RDBMS). The sequences were clustered using BLAST with a product score of 50. All clusters of two or more sequences created a bin which represents one transcribed gene.

Assembly of the component sequences within each bin was performed using a modification of Phrap, a publicly available program for assembling DNA fragments (Green, P. University of Washington, Seattle WA). Bins that showed 82% identity from a local pair-wise alignment between any of the

PB-0017 US

consensus sequences were merged.

Bins were annotated by screening the consensus sequence in each bin against public databases, such as GBpri and GenPept from NCBI. The annotation process involved a FASTn screen against the GBpri database in GenBank. Those hits with a percent identity of greater than or equal to 75% and an alignment length of greater than or equal to 100 base pairs were recorded as homolog hits. The residual unannotated sequences were screened by FASTx against GenPept. Those hits with an E value of less than or equal to 10^{-8} were recorded as homolog hits.

Sequences were then reclustered using BLASTn and Cross-Match, a program for rapid amino acid and nucleic acid sequence comparison and database search (Green, supra), sequentially. Any BLAST alignment between a sequence and a consensus sequence with a score greater than 150 was realigned using cross-match. The sequence was added to the bin whose consensus sequence gave the highest Smith-Waterman score (Smith et al. (1992) Protein Engineering 5:35-51) amongst local alignments with at least 82% identity. Non-matching sequences were moved into new bins, and assembly processes were repeated.

IV Library Subtraction

A profile of the polynucleotides that reflect gene transcription activity in a particular tissue at a particular time is defined as a "transcript image". Such profiles are produced by sequencing cDNAs and then naming, matching, and counting all copies of related clones and arranging them in order of abundance. The process of producing a comparative transcript image was fully described in USPN 5,840,484, incorporated herein by reference.

Subtractions between transcript images show the differences occurring between cDNAs produced in untreated and retinoic acid treated fibroblast and bone marrow libraries. The FIBRUNT01 library contained 3781 cDNA clones, and the FIBRTXT02 library, 3849; the BMARUNT02 library contained 3715 cDNA clones, and the BMARTXT02, 3375.

Large numbers of mRNA transcripts, as represented by their respective cDNA clones, were compared using computer-based or "electronic subtraction" methods. For purposes of example, electronic subtraction between any two transcript images parallels hybrid subtraction between any two cDNA libraries using techniques that are known to those of skill in the art (Meyers, supra, pp. 698-699).

V Description of Known Genes**Retinoic Acid Induced Genes**

<u>Name</u>	<u>Description</u>
RAI-1/2	RAI-1 and -2 are retinoic acid-metabolizing cytochromes which regulate retinoic acid metabolism. RAI-2 is predominantly expressed in the adult cerebellum and is responsible for all-trans-retinoic acid metabolism. RAI-1 is a novel polyglutamine encoding gene that is deleted in Smith-Magenis syndrome patients. (White <u>et al.</u> (1996) J Biol Chem 271:29922-

PB-0017 US

29927; White *et al.* (2000) *Proc Natl Acad Sci* 97:6403-6408; and Loudig *et al.* (2000) *Mol Endocrinol* 14:1483-1497).

CYP26

Retinoic acid hydroxylase (CYP26), is highly specific for all-trans-RA, can be induced through RA receptors in human breast and colon carcinoma cells, and is a key enzyme in neuronal differentiation of embryonal carcinoma cells (Marikar *et al.* (1998) *J Invest Dermatol* 111:434-439; Sonneveld *et al.* (1999) *Dev Biol* 213:390-404; and Sonneveld *et al.* (1998) *Cell Growth Differ* 9:629-637).

B-2 laminin
is

Beta 2 laminin, is a component of the extracellular matrix that affects retinal development, lacking in congenital muscular dystrophy, and has been implicated in cell growth inhibition and differentiation in testicular carcinoma cells (Bielinska and Wilson (1997) *Mech Dev* 65:43-54; Clifford *et al.* (1999) *Cancer Res* 59:14-18; and Ueno *et al.* (1997) *Hum Cell* 10:151-158).

RBPR-p63

In one form, retinol-binding-protein receptor protein 63 (RBPR-p63) is an integral membrane protein of the retinal epithelium where it is part of the receptor-binding complex that mediates the uptake RBP-bound vitamin A. In another form, it acts with other genes to transactivate the promoter of keratinocyte differentiation genes (Bavik *et al.* (1993) *J Biol Chem* 268:20540-20546; Johansson *et al.* (1997) *Anat Embryol* 195:483-490; De Laurenzi *et al.* (2000) *Biochem Biophys Res Commun* 273:342-346 and Levrero *et al.* (2000) *J Cell Sci* 113:1661-1670).

P97

The tyrosine kinase phosphorylation of the P97 is associated with different cellular activities and necessary for cell free constitution of the transitional ER. It has been described as a translational regulator, implicated in the induction of Myc-intron-binding polypeptides, MIBP1 and RFX1, during retinoic acid-mediated differentiation of haemopoietic cells, and associated with allergic asthma response (Imataka *et al.* (1997) *EMBO J* 16:817-825; Zajack-Kaye *et al.* (2000) *Biochem J* 345:535-541).

lamin A

Lamin A is found in an unpolymerized state throughout the nucleoplasm of daughter cell nuclei in early G1 and gradually becomes incorporated into the peripheral lamina during the first few hours of this stage of the cell cycle. It is induced *in vitro* during differentiation of F9 and P19 embryonal carcinoma cells (Lebel and Raymond (1987) *Biochem Biophys Res Commun* 149:417-423; Mattia *et al.* (1992) *Exp Cell Res* 203:449-455; and Okumura *et al.* (2000) *Biochem Biophys Res Commun* 269:197-202).

 β -2-m

Beta-2-microglobulin (β -2-m) is an amyloid protein implicated in the retinoic acid-induced differentiation of F9, neuroblastoma cells, and myeloma tumors. It is one of the proteins diagnostic of segmental glomerulosclerosis, contributing to dialysis-related amyloidosis and specifically removed during hemodialysis (Eriksson *et al.* (1986) *Cancer Res* 46:717-722; Hanada *et al.* (1993) *Cancer Res* 53:4978-4986; Lonergan *et al.* (1993) *Mol Cell Biol* 13:6629-6639, and Segars *et al.* (1993) *Mol Cell Biol* 13:6157-6169).

APLP2

The amyloid precursor like protein 2 (APLP2), implicated in Alzheimer's disease, exhibits an intrinsic affinity for the major histocompatibility complex K(d) molecule and strictly depends upon the presence of β -2-m. APLP2 gene expression is increased in human neuroblastoma cells in response to retinoic acid (Beckman and Iverfeldt (1997) *Neurosci Lett* 221:73-76).

PB-0017 US

Genes Involved in Cell Signaling and Differentiation

Name	Description
Collagens	Collagen distribution is altered in the basement membranes of various adenocarcinomas, in the serum of patients with metastatic breast cancer and collagen vascular diseases and in intra-alveolar spaces of patients with interstitial lung disease (Maemura <u>et al.</u> (2000) <i>Oncol Rep</i> 7:1333-1338; Amenta <u>et al.</u> (2000) <i>Hum Pathol</i> 31:359-366; Horie <u>et al.</u> (2000) <i>J Rheumatol</i> 27:2378-2381; and Yasui <u>et al.</u> (2000) <i>Clin Appl Thromb Hemost</i> 6:202-205).
Notch 3	The Notch signaling genes are pivotal for cell fate decisions at many stages of development and particularly during formation of the nervous system. Notch 3 lacks specific epidermal growth factor repeats and is expressed in proliferating neuroepithelium, during pancreatic development, and in hematopoiesis (Mitsiadis <u>et al.</u> (1998) <i>Dev Biol</i> 204:420-431; Apelqvist <u>et al.</u> (1999) <i>Nature</i> 400:877-881; and Singh <u>et al.</u> (2000) <i>Exp Hematol</i> 28:527-534).
PDGF	Platelet derived growth factor is expressed during embryogenesis, specifically retinoic acid stimulates immature lung fibroblast growth via a PDGF-mediated autocrine mechanism and plays a role in the regulation of oligodendrocyte differentiation in the spinal cord (Baron <u>et al.</u> (2000) <i>Mol Cell Neurosci</i> 15:314-329; Liebeskind <u>et al.</u> (2000) <i>Am J Physiol Lung Cell Mol Physiol</i> 279:L81-90; Yen and Varvayanis (2000) <i>In Vitro Cell Dev Biol Anim</i> 36:249-255; and Noll and Miller (1994) <i>Development</i> 120:649-660).
Fibulin-2	The extracellular matrix protein, fibulin-2, participates in embryogenesis including atrioventricular valvuloseptal morphogenesis, and is a stromal component of tumors (Eisenberg and Markwald (1995) <i>Circ Res</i> 77:1-6; Miosge <u>et al.</u> (1996) <i>Histochem J</i> 28:109-116).
Fibrillin	Retinoic acid affects the EGF-R signaling pathway during differentiation of human endometrial adenocarcinoma cells and during expression of early markers of precardiac asymmetry (Eisenberg and Markwald (1995) <i>Circ Res</i> 77:1-6; Smith <u>et al.</u> (1997) <i>Dev Biol</i> 182:162-171).
IGF-I	Expression of insulin-like growth factor I (IGF-I) is regulated by retinoic acid in osteoblasts. (Gabbitas and Canalis (1997) <i>Cell Physiol</i> 172:253-264).
PKR RI	cAMP-dependent protein kinase regulatory subunit RI beta participates in retinoic acid induced differentiation; in fact, retinoylation of the protein is increased in psoriatic human fibroblasts (Tournier <u>et al.</u> (1996) <i>J Cell Physiol</i> 167:196-203).
SWAP	SWAP is a regulatory gene under hormonal control during neuronal differentiation.
EMP	EMP is a protein that promotes terminal differentiation of erythroblasts by suppressing apoptosis (Hanspal <u>et al.</u> (1998) <i>Blood</i> 92:2940-2950).
Rab 11	YPT3/rab11 is a member of the ras family of signaling molecules. Expression of rab GTP binding proteins has been implicated in oligodendrocyte differentiation, and altered growth regulation and loss of response to retinoic acid accompany tumorigenic transformation of prostatic cells (Drivas <u>et al.</u> (1991) <i>Oncogene</i> 6:3-9; Lai <u>et al.</u> (1994) <i>Genomics</i> 22:610-616; Peehl <u>et al.</u> (1999) <i>Anticancer Res</i> 19:3857-3864; and Bouverat <u>et al.</u> (2000) <i>J Neurosci Res</i> 59:446-453).

PB-0017 US

ERK-1 ERK-1 is a member of the Ras-extracellular signal-regulated kinase signaling pathway involved in brain-derived neurotrophic factor-mediated survival and neurogenesis, in the induction of erythroid differentiation, and various other cellular differentiation processes (Encinas *et al.* (1999) *J Neurochem* 73:1409-1421; Matsuzaki *et al.* (2000) *Oncogene* 19:1500-1508; Nguyen *et al.* (2000) *J Biol Chem* 275:19382-19388).

β4 integrin Retinoic acid advances expression of β4 integrin which in turn affects development of the embryonic heart (Hierck *et al.* (1996) *Dev Dyn* 207:89-103; Laurikainen *et al.* (1996) *Arch Dermatol Res* 288:270-273).

VI Co-Expression Among the Known Genes and Novel cDNAs

Using the LIFESEQ GOLD database (Incyte Genomics), five cDNAs that showed significant association with known retinoic acid induced genes were identified. Initially, degree of association between known the known retinoic acid induced genes and cDNAs were measured by probability values using a cutoff p-value less than 0.00001. The process was reiterated so that an initial selection of genes were reduced to the final five cDNAs claimed. The following tabular entries show the p-value for the co-expression of any two genes compared with cell differentiation genes. The cDNAs are identified by their SEQ ID NOs, and the known genes, by their abbreviations as shown herein. For each cDNA, the p-value is the probability that the observed co-expression is due to chance, using the Fisher Exact Test.

Genes co-expressed with SEQ ID NO:1	p-value
Collagen type I	6.79e-16
Notch3	6.71e-15
PDGF	4.73e-14
Fibulin-2	4.85e-12
Fibrillin	1.79e-11
IGF-I	1.18e-10
Collagen type XV alpha-1	1.69e-10
Genes co-expressed with SEQ ID NO:2	p-value
PKR RI	3.86e-12
SWAP	4.76e-12
Genes co-expressed with SEQ ID NO:3	p-value
YPT3/rab11	1.24e-13
ERK-1	1.83e-13
β 4 integrin	3.68e-13
Genes co-expressed with SEQ ID NO:4	p-value
EMP	4.89e-14

Of the five genes, only SEQ ID NO:5 has homology with a known gene, the integrase interactor 1α protein. The cDNA of SEQ ID NO:5 encodes a member of the SWI/SNF chromatin remodeling

PB-0017 US

complex that regulates transcription factor access to regulatory DNA sequences and has been identified as a tumor suppressor for rhabdoid tumors (Suzuki *et al.* (1997) *Diagn Mol Pathol* 6:326-332; Biegel *et al.* (1999) *Cancer Res* 59:74-79; DeCristofaro *et al.* (1999) *Oncogene* 18: 7559-7565; Sevenet *et al.* (1999) *Hum Mol Genet* 8:2359-2368; Sevenet *et al.* (1999) *Am J Hum Genet* 65:1342-1348; and Manda *et al.* (2000) *Cancer Lett* 153:57-61). Upregulation of this gene is consistent with the role of retinoic acid as an anti-tumor agent.

VII Homology Searching of cDNA Clones and Their Deduced Proteins

The cDNAs of the Sequence Listing or their deduced amino acid sequences were used to query databases such as GenBank, SwissProt, BLOCKS, and the like. These databases that contain previously identified and annotated sequences or domains were searched using BLAST or BLAST 2 (Altschul *et al.* *supra*; Altschul, *supra*) to produce alignments and to determine which sequences were exact matches or homologs. The alignments were to sequences of prokaryotic (bacterial) or eukaryotic (animal, fungal, or plant) origin. Alternatively, algorithms such as the one described in Smith and Smith (1992, *Protein Engineering* 5:35-51) could have been used to deal with primary sequence patterns and secondary structure gap penalties. All of the sequences disclosed in this application have lengths of at least 49 nucleotides, and no more than 12% uncalled bases (where N is recorded rather than A, C, G, or T).

As detailed in Karlin (*supra*), BLAST matches between a query sequence and a database sequence were evaluated statistically and only reported when they satisfied the threshold of 10^{-25} for nucleotides and 10^{-14} for peptides. Homology was also evaluated by product score calculated as follows: the % nucleotide or amino acid identity [between the query and reference sequences] in BLAST is multiplied by the % maximum possible BLAST score [based on the lengths of query and reference sequences] and then divided by 100. In comparison with hybridization procedures used in the laboratory, the electronic stringency for an exact match was set at 70, and the conservative lower limit for an exact match was set at approximately 40 (with 1-2% error due to uncalled bases).

The BLAST software suite, freely available sequence comparison algorithms (NCBI, Bethesda MD), includes various sequence analysis programs including "blastn" that is used to align nucleic acid molecules and BLAST 2 that is used for direct pairwise comparison of either nucleic or amino acid molecules. BLAST programs are commonly used with gap and other parameters set to default settings, e.g.: Matrix: BLOSUM62; Reward for match: 1; Penalty for mismatch: -2; Open Gap: 5 and Extension Gap: 2 penalties; Gap x drop-off: 50; Expect: 10; Word Size: 11; and Filter: on. Identity or similarity is measured over the entire length of a sequence or some smaller portion thereof. Brenner *et al.* (1998; *Proc Natl Acad Sci* 95:6073-6078, incorporated herein by reference) analyzed the BLAST for its ability to identify structural homologs by sequence identity and found 30% identity is a reliable threshold for sequence alignments of at least 150 residues and 40%, for alignments of at least 70 residues.

PB-0017 US

The cDNAs of this application were compared with assembled consensus sequences or templates found in the LIFESEQ GOLD database. Component sequences from cDNA, extension, full length, and shotgun sequencing projects were subjected to PHRED analysis and assigned a quality score. All sequences with an acceptable quality score were subjected to various pre-processing and editing pathways to remove low quality 3' ends, vector and linker sequences, polyA tails, Alu repeats, mitochondrial and ribosomal sequences, and bacterial contamination sequences. Edited sequences had to be at least 50 bp in length, and low-information sequences and repetitive elements such as dinucleotide repeats, Alu repeats, and the like, were replaced by "Ns" or masked.

Edited sequences were subjected to assembly procedures in which the sequences were assigned to gene bins. Each sequence could only belong to one bin, and sequences in each bin were assembled to produce a template. Newly sequenced components were added to existing bins using BLAST and CROSSMATCH. To be added to a bin, the component sequences had to have a BLAST quality score greater than or equal to 150 and an alignment of at least 82% local identity. The sequences in each bin were assembled using PHRAP. Bins with several overlapping component sequences were assembled using DEEP PHRAP. The orientation of each template was determined based on the number and orientation of its component sequences.

Bins were compared to one another and those having local similarity of at least 82% were combined and reassembled. Bins having templates with less than 95% local identity were split. Templates were subjected to analysis by STITCHER/EXON MAPPER algorithms that analyze the probabilities of the presence of splice variants, alternatively spliced exons, splice junctions, differential expression of alternative spliced genes across tissue types or disease states, and the like. Assembly procedures were repeated periodically, and templates were annotated using BLAST against GenBank databases such as GBpri. An exact match was defined as having from 95% local identity over 200 base pairs through 100% local identity over 100 base pairs and a homolog match as having an E-value (or probability score) of $\leq 1 \times 10^{-8}$. The templates were also subjected to frameshift FASTx against GENPEPT, and homolog match was defined as having an E-value of $\leq 1 \times 10^{-8}$. Template analysis and assembly was described in USSN 09/276,534, filed March 25, 1999.

Following assembly, templates were subjected to BLAST, motif, and other functional analyses and categorized in protein hierarchies using methods described in USSN 08/812,290 and USSN 08/811,758, both filed March 6, 1997; in USSN 08/947,845, filed October 9, 1997; and in USSN 09/034,807, filed March 4, 1998. Then templates were analyzed by translating each template in all three forward reading frames and searching each translation against the PFAM database of hidden Markov model-based protein families and domains using the HMMER software package (Washington University School of Medicine, St. Louis MO; <http://pfam.wustl.edu/>).

PB-0017 US

The cDNA was further analyzed using MACDNASIS PRO software (Hitachi Software Engineering), and LASERGENE software (DNASTAR) and queried against public databases such as the GenBank rodent, mammalian, vertebrate, prokaryote, and eukaryote databases, SwissProt, BLOCKS, PRINTS, PFAM, and Prosite.

VIII Hybridization Technologies and Analyses

Immobilization of cDNAs on a Substrate

The cDNAs are applied to a substrate by one of the following methods. A mixture of cDNAs is fractionated by gel electrophoresis and transferred to a nylon membrane by capillary transfer. Alternatively, the cDNAs are individually ligated to a vector and inserted into bacterial host cells to form a library. The cDNAs are then arranged on a substrate by one of the following methods. In the first method, bacterial cells containing individual clones are robotically picked and arranged on a nylon membrane. The membrane is placed on LB agar containing selective agent (carbenicillin, kanamycin, ampicillin, or chloramphenicol depending on the vector used) and incubated at 37C for 16 hr. The membrane is removed from the agar and consecutively placed colony side up in 10% SDS, denaturing solution (1.5 M NaCl, 0.5 M NaOH), neutralizing solution (1.5 M NaCl, 1 M Tris-HCl, pH 8.0), and twice in 2xSSC for 10 min each. The membrane is then UV irradiated in a STRATALINKER UV-crosslinker (Stratagene).

In the second method, cDNAs are amplified from bacterial vectors by thirty cycles of PCR using primers complementary to vector sequences flanking the insert. PCR amplification increases a starting concentration of 1-2 ng nucleic acid to a final quantity greater than 5 μ g. Amplified nucleic acids from about 400 bp to about 5000 bp in length are purified using SEPHACRYL-400 beads (APB). Purified nucleic acids are arranged on a nylon membrane manually or using a dot/slot blotting manifold and suction device and are immobilized by denaturation, neutralization, and UV irradiation as described above. Purified nucleic acids are robotically arranged and immobilized on polymer-coated glass slides using the procedure described in USPN 5,807,522. Polymer-coated slides are prepared by cleaning glass microscope slides (Corning, Acton MA) by ultrasound in 0.1% SDS and acetone, etching in 4% hydrofluoric acid (VWR Scientific Products, West Chester PA), coating with 0.05% aminopropyl silane (Sigma-Aldrich) in 95% ethanol, and curing in a 110C oven. The slides are washed extensively with distilled water between and after treatments. The nucleic acids are arranged on the slide and then immobilized by exposing the array to UV irradiation using a STRATALINKER UV-crosslinker (Stratagene). Arrays are then washed at room temperature in 0.2% SDS and rinsed three times in distilled water. Non-specific binding sites are blocked by incubation of arrays in 0.2% casein in phosphate buffered saline (PBS; Tropix, Bedford MA) for 30 min at 60C; then the arrays are washed in 0.2% SDS and rinsed in distilled water as before.

PB-0017 US

Probe Preparation for Membrane Hybridization

Hybridization probes derived from the cDNAs of the Sequence Listing are employed for screening cDNAs, mRNAs, or genomic DNA in membrane-based hybridizations. Probes are prepared by diluting the cDNAs to a concentration of 40-50 ng in 45 μ l TE buffer, denaturing by heating to 100C for five min, and briefly centrifuging. The denatured cDNA is then added to a REDIPRIME tube (APB), gently mixed until blue color is evenly distributed, and briefly centrifuged. Five μ l of [³²P]dCTP is added to the tube, and the contents are incubated at 37C for 10 min. The labeling reaction is stopped by adding 5 μ l of 0.2M EDTA, and probe is purified from unincorporated nucleotides using a PROBEQUANT G-50 microcolumn (APB). The purified probe is heated to 100C for five min, snap cooled for two min on ice, and used in membrane-based hybridizations as described below.

Probe Preparation for Polymer Coated Slide Hybridization

Hybridization probes derived from mRNA isolated from samples are employed for screening cDNAs of the Sequence Listing in array-based hybridizations. Probe is prepared using the GEMbright kit (Incyte Genomics) by diluting mRNA to a concentration of 200 ng in 9 μ l TE buffer and adding 5 μ l 5x buffer, 1 μ l 0.1 M DTT, 3 μ l Cy3 or Cy5 labeling mix, 1 μ l RNase inhibitor, 1 μ l reverse transcriptase, and 5 μ l 1x yeast control mRNAs. Yeast control mRNAs are synthesized by *in vitro* transcription from noncoding yeast genomic DNA (W. Lei, unpublished). As quantitative controls, one set of control mRNAs at 0.002 ng, 0.02 ng, 0.2 ng, and 2 ng are diluted into reverse transcription reaction mixture at ratios of 1:100,000, 1:10,000, 1:1000, and 1:100 (w/w) to sample mRNA respectively. To examine mRNA differential expression patterns, a second set of control mRNAs are diluted into reverse transcription reaction mixture at ratios of 1:3, 3:1, 1:10, 10:1, 1:25, and 25:1 (w/w). The reaction mixture is mixed and incubated at 37C for two hr. The reaction mixture is then incubated for 20 min at 85C, and probes are purified using two successive CHROMA SPIN+TE 30 columns (Clontech). Purified probe is ethanol precipitated by diluting probe to 90 μ l in DEPC-treated water, adding 2 μ l 1mg/ml glycogen, 60 μ l 5 M sodium acetate, and 300 μ l 100% ethanol. The probe is centrifuged for 20 min at 20,800xg, and the pellet is resuspended in 12 μ l resuspension buffer, heated to 65C for five min, and mixed thoroughly. The probe is heated and mixed as before and then stored on ice. Probe is used in high density array-based hybridizations as described below.

Membrane-based Hybridization

Membranes are pre-hybridized in hybridization solution containing 1% Sarkosyl and 1x high phosphate buffer (0.5 M NaCl, 0.1 M Na₂HPO₄, 5 mM EDTA, pH 7) at 55C for two hr. The probe, diluted in 15 ml fresh hybridization solution, is then added to the membrane. The membrane is hybridized with the probe at 55C for 16 hr. Following hybridization, the membrane is washed for 15 min at 25C in 1mM Tris (pH 8.0), 1% Sarkosyl, and four times for 15 min each at 25C in 1mM Tris (pH 8.0). To detect

PB-0017 US

hybridization complexes, XOMAT-AR film (Eastman Kodak, Rochester NY) is exposed to the membrane overnight at

-70C, developed, and examined visually.

5 Polymer Coated Slide-based Hybridization

Probe is heated to 65C for five min, centrifuged five min at 9400 rpm in a 5415C microcentrifuge (Eppendorf Scientific, Westbury NY), and then 18 μ l are aliquoted onto the array surface and covered with a coverslip. The arrays are transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber is kept at 100% humidity internally by the addition of 140 μ l of 5xSSC in a corner of the chamber. The chamber containing the arrays is incubated for about 6.5 hr at 60C. The arrays are washed for 10 min at 45C in 1xSSC, 0.1% SDS, and three times for 10 min each at 45C in 0.1xSSC, and dried.

Hybridization reactions are performed in absolute or differential hybridization formats. In the absolute hybridization format, probe from one sample is hybridized to array elements, and signals are detected after hybridization complexes form. Signal strength correlates with probe mRNA levels in the sample. In the differential hybridization format, differential expression of a set of genes in two biological samples is analyzed. Probes from the two samples are prepared and labeled with different labeling moieties. A mixture of the two labeled probes is hybridized to the array elements, and signals are examined under conditions in which the emissions from the two different labels are individually detectable. Elements on the array that are hybridized to substantially equal numbers of probes derived from both biological samples give a distinct combined fluorescence (Shalon WO95/35505).

Hybridization complexes are detected with a microscope equipped with an INNOVA 70 mixed gas 10 W laser (Coherent, Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light is focused on the array using a 20X microscope objective (Nikon, Melville NY). The slide containing the array is placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective with a resolution of 20 micrometers. In the differential hybridization format, the two fluorophores are sequentially excited by the laser. Emitted light is split, based on wavelength, into two photomultiplier tube detectors (PMT R1477, Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Appropriate filters positioned between the array and the photomultiplier tubes are used to filter the signals. The emission maxima of the fluorophores used are 565 nm for Cy3 and 650 nm for Cy5. The sensitivity of the scans is calibrated using the signal intensity generated by the yeast control mRNAs added to the probe mix. A specific location on the array contains a complementary DNA sequence, allowing the intensity of the signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000.

PB-0017 US

The output of the photomultiplier tube is digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Norwood MA) installed in an IBM-compatible PC computer. The digitized data are displayed as an image where the signal intensity is mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data is also analyzed quantitatively. Where two different fluorophores are excited and measured simultaneously, the data are first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using the emission spectrum for each fluorophore. A grid is superimposed over the fluorescence signal image such that the signal from each spot is centered in each element of the grid. The fluorescence signal within each element is then integrated to obtain a numerical value corresponding to the average intensity of the signal. The software used for signal analysis is the GEMTOOLS program (Incyte Genomics).

IX Expression Analysis

BLAST was used to search for identical or related molecules in the GenBank or LIFESEQ databases (Incyte Genomics). The product score for human and rat sequences was calculated as follows: the BLAST score is multiplied by the % nucleotide identity and the product is divided by (5 times the length of the shorter of the two sequences), such that a 100% alignment over the length of the shorter sequence gives a product score of 100. The product score takes into account both the degree of similarity between two sequences and the length of the sequence match. For example, with a product score of 40, the match will be exact within a 1% to 2% error, and with a product score of at least 70, the match will be exact. Similar or related molecules are usually identified by selecting those which show product scores between 8 and 40.

Electronic northern analysis was performed at a product score of 70 are shown in Fig. 4. All sequences and cDNA libraries in the LIFESEQ database were categorized by system, organ/tissue and cell type. The categories included cardiovascular system, connective tissue, digestive system, embryonic structures, endocrine system, exocrine glands, female and male genitalia, germ cells, hemic/immune system, liver, musculoskeletal system, nervous system, pancreas, respiratory system, sense organs, skin, stomatognathic system, unclassified/mixed, and the urinary tract. For each category, the number of libraries in which the sequence was expressed were counted and shown over the total number of libraries in that category. In a non-normalized library, expression levels of two or more are significant.

X Complementary Molecules

Molecules complementary to the cDNA, from about 5 (PNA) to about 5000 bp (complement of a cDNA insert), are used to detect or inhibit gene expression. These molecules are selected using LASERGENE software (DNASTAR). Detection is described in Example VII. To inhibit transcription by preventing promoter binding, the complementary molecule is designed to bind to the most unique 5'

PB-0017 US

sequence and includes nucleotides of the 5' UTR upstream of the initiation codon of the open reading frame. Complementary molecules include genomic sequences (such as enhancers or introns) and are used in "triple helix" base pairing to compromise the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. To inhibit translation, a complementary molecule is designed to prevent ribosomal binding to the mRNA encoding the protein.

Complementary molecules are placed in expression vectors and used to transform a cell line to test efficacy; into an organ, tumor, synovial cavity, or the vascular system for transient or short term therapy; or into a stem cell, zygote, or other reproducing lineage for long term or stable gene therapy. Transient expression lasts for a month or more with a non-replicating vector and for three months or more if appropriate elements for inducing vector replication are used in the transformation/expression system.

Stable transformation of appropriate dividing cells with a vector encoding the complementary molecule produces a transgenic cell line, tissue, or organism (USPN 4,736,866). Those cells that assimilate and replicate sufficient quantities of the vector to allow stable integration also produce enough complementary molecules to compromise or entirely eliminate activity of the cDNA encoding the protein.

XI Protein Expression

Expression and purification of the protein are achieved using either a cell expression system or an insect cell expression system. The pUB6/V5-His vector system (Invitrogen) is used to express protein in CHO cells. The vector contains the selectable bsd gene, multiple cloning sites, the promoter/enhancer sequence from the human ubiquitin C gene, a C-terminal V5 epitope for antibody detection with anti-V5 antibodies, and a C-terminal polyhistidine (6xHis) sequence for rapid purification on PROBOND resin (Invitrogen). Transformed cells are selected on media containing blasticidin.

Spodoptera frugiperda (Sf9) insect cells are infected with recombinant Autographica californica nuclear polyhedrosis virus (baculovirus). The polyhedrin gene is replaced with the cDNA by homologous recombination and the polyhedrin promoter drives cDNA transcription. The protein is synthesized as a fusion protein with 6xhis which enables purification as described above. Purified protein is used in the following activity and to make antibodies

XII Production of Antibodies

The protein is purified using polyacrylamide gel electrophoresis and used to immunize mice or rabbits. Antibodies are produced using the protocols below. Alternatively, the amino acid sequence of the expressed protein is analyzed using LASERGENE software (DNASTAR) to determine regions of high antigenicity. An antigenic epitope, usually found near the C-terminus or in a hydrophilic region is selected, synthesized, and used to raise antibodies. Typically, epitopes of about 15 residues in length are produced using an ABI 431A peptide synthesizer (ABI) using FMOC-chemistry and coupled to KLH (Sigma-Aldrich) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester to increase antigenicity.

PB-0017 US

Rabbits are immunized with the epitope-KLH complex in complete Freund's adjuvant. Immunizations are repeated at intervals thereafter in incomplete Freund's adjuvant. After a minimum of seven weeks for mouse or twelve weeks for rabbit, antisera are drawn and tested for antipeptide activity. Testing involves binding the peptide to plastic, blocking with 1% bovine serum albumin, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG. Methods well known in the art are used to determine antibody titer and the amount of complex formation.

XIII Purification of Naturally Occurring Protein Using Specific Antibodies

Naturally occurring or recombinant protein is purified by immunoaffinity chromatography using antibodies which specifically bind the protein. An immunoaffinity column is constructed by covalently coupling the antibody to CNBr-activated SEPHAROSE resin (APB). Media containing the protein is passed over the immunoaffinity column, and the column is washed using high ionic strength buffers in the presence of detergent to allow preferential absorbance of the protein. After coupling, the protein is eluted from the column using a buffer of pH 2-3 or a high concentration of urea or thiocyanate ion to disrupt antibody/protein binding, and the protein is collected.

XIV Screening Molecules for Specific Binding with the cDNA or Protein

The cDNA, or fragments thereof, or the protein, or portions thereof, are labeled with ^{32}P -dCTP, Cy3-dCTP, or Cy5-dCTP (APB), or with BIODIPY or FITC (Molecular Probes, Eugene OR), respectively. Libraries of candidate molecules or compounds previously arranged on a substrate are incubated in the presence of labeled cDNA or protein. After incubation under conditions for either a nucleic acid or amino acid sequence, the substrate is washed, and any position on the substrate retaining label, which indicates specific binding or complex formation, is assayed, and the ligand is identified. Data obtained using different concentrations of the nucleic acid or protein are used to calculate affinity between the labeled nucleic acid or protein and the bound molecule.

XV Two-Hybrid Screen

A yeast two-hybrid system, MATCHMAKER LexA Two-Hybrid system (Clontech Laboratories), is used to screen for peptides that bind the protein of the invention. A cDNA encoding the protein is inserted into the multiple cloning site of a pLexA vector, ligated, and transformed into *E. coli*. cDNA, prepared from mRNA, is inserted into the multiple cloning site of a pB42AD vector, ligated, and transformed into *E. coli* to construct a cDNA library. The pLexA plasmid and pB42AD-cDNA library constructs are isolated from *E. coli* and used in a 2:1 ratio to co-transform competent yeast EGY48[p8op-lacZ] cells using a polyethylene glycol/lithium acetate protocol. Transformed yeast cells are plated on synthetic dropout (SD) media lacking histidine (-His), tryptophan (-Trp), and uracil (-Ura), and incubated at 30C until the colonies have grown up and are counted. The colonies are pooled in a minimal volume of 1x TE (pH 7.5), replated on SD/-His/-Leu/-Trp/-Ura media supplemented with 2% galactose (Gal), 1%

PB-0017 US

raffinose (Raf), and 80 mg/ml 5-bromo-4-chloro-3-indolyl β -d-galactopyranoside (X-Gal), and subsequently examined for growth of blue colonies. Interaction between expressed protein and cDNA fusion proteins activates expression of a LEU2 reporter gene in EGY48 and produces colony growth on media lacking leucine (-Leu). Interaction also activates expression of β -galactosidase from the p8op-lacZ reporter construct that produces blue color in colonies grown on X-Gal.

Positive interactions between expressed protein and cDNA fusion proteins are verified by isolating individual positive colonies and growing them in SD/-Trp/-Ura liquid medium for 1 to 2 days at 30C. A sample of the culture is plated on SD/-Trp/-Ura media and incubated at 30C until colonies appear. The sample is replica-plated on SD/-Trp/-Ura and SD/-His/-Trp/-Ura plates. Colonies that grow on SD containing histidine but not on media lacking histidine have lost the pLexA plasmid. Histidine-requiring colonies are grown on SD/Gal/Raf/X-Gal/-Trp/-Ura, and white colonies are isolated and propagated. The pB42AD-cDNA plasmid, which contains a cDNA encoding a protein that physically interacts with the protein, is isolated from the yeast cells and characterized.

All patents and publications mentioned in the specification are incorporated by reference herein. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.